Doctoral Programme in Sustainable Use of Renewable Natural Resources
Department of Agricultural Sciences
Faculty of Agriculture and Forestry (Animal Science)
University of Helsinki
Helsinki

SINGLE-STEP GENOMIC PREDICTION IN SMALL-SCALE POPULATIONS

DOCTORAL THESIS

Kudinov A. Andrei

ACADEMIC DISSERTATION

To be presented for public examination with the permission of the Faculty of Agriculture and Forestry of the University of Helsinki, in lecture room 115, Fabianinkatu 26, Helsinki, on 17th June 2021, at 14 o'clock.

Helsinki 2021

Custos: Professor Pekka Uimari
University of Helsinki, Finland

Supervisors: Research Professor Ismo Strandén
Natural Resources Institute Finland

Professor Esa Mäntysaari
Natural Resources Institute Finland

Professor Pekka Uimari
University of Helsinki, Finland

Co-supervisor: Professor Kirill Plemyashov
Saint-Petersburg State University
of Veterinary Medicine, Russia

Reviewers: Associate Professor Mario Calus
Wageningen University and Research, The Netherlands

Professor Raphael Mrode
Scotland's Rural College, United Kingdom

Opponent: Senior Researcher Ole Christensen
Center of Quantitative Genetics and Genomics,
Aarhus University, Denmark

Dissertationes Schola Doctoralis Scientiae Circumiectalis,
Alimentariae, Biologicae (9/2021)
ISBN 978-951-51-7296-9 (Print)
ISBN 978-951-51-7297-6 (Online)
ISSN 2342-5423 (Print)
ISSN 2342-5431 (Online)

Electronic publication at https://ethesis.helsinki.fi

The Faculty of Agriculture and Forestry of the University of Helsinki uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

Unigrafia
Helsinki 2021

“Before enlightenment: chop wood, carry water.
After enlightenment: chop wood, carry water.”
Zen Kōan

Abstract

Selection based on genomic enhanced breeding values (GEBVs) has been successful in dairy cattle breeding over the last decade. A set of reference animals can be used to estimate genomic marker effect solutions which are further used to compute the GEBVs of candidate animals. Implementation of genomic evaluation is challenging for populations with limited data. The main objective of this thesis was to identify an approach to implement single-step genomic best linear unbiased prediction (ssGBLUP) in a small-scale dairy cattle population. In particular, the aims were to predict GEBVs by ssGBLUP using local genotypes, enhance the reliability of prediction by incorporating data from an external breeding population, and implement the metafounder approach.

The first objective was to develop breeding value evaluation for the Leningrad region (LR) in Russia. For many years, the LR has been the leading national dairy region according to average milk production per cow (8,681 kg in 2017). Despite farmer interest in obtain GEBVs for young animals, implementation of a genomic prediction model over the official dairy evaluation method (Contemporary Comparison) has not been realized. Therefore, the first objective of this study was to develop state-of-the-art BLUP Animal Model for Holstein (HOL) and Russian Black & White cattle (Publication I). The traits of focus were milk, fat, and protein yield. The data used to develop the first (FLM) and multiple (MLM) lactation models included 320,633 repeated 305d records from 49 herds. The heritability estimates for milk, fat, and protein yield were 0.24, 0.20, and 0.20 for FLM and 0.18, 0.19, and 0.18 for MLM. For cows born between 2000 and 2016, MLM estimated an annual average genetic gain of 56 kg, 1.84 kg, and 1.62 kg for milk, fat, and protein yield, respectively.

The second objective was to implement ssGBLUP for LR using a set of local genotyped animals (Publ. II). Genomic data were available from 1080 cows and 427 bulls. MLM was improved by adding a herd by sire interaction random effect. The traits used were milk and fat yield with the same heritability estimates: 0.21. Milk yield cross-validation analysis showed a validation reliability (R^2) of 0.21 and 0.38 for bulls and cows, respectively. The R^2 values for fat yield for bulls and cows were 0.17 and 0.41.

The third objective was to enhance the LR ssGBLUP prediction by using external DFS (Denmark, Finland, and Sweden) HOL genomic and pedigree information (Publ. II). Data included 414 bull genotypes and 487 milk and fat EBVs published by Interbull on the DFS scale. The inclusion of DFS genotypes did not change the milk yield R^2 for bulls but slightly decreased it for cows (0.38 to 0.36). For fat yield, R^2 increased from 0.17 to 0.18 for bulls and decreased from 0.41 to 0.34 for cows. In analysis of milk yield, the highest R^2 was realized in the ssGBLUP model simultaneously using genomic and phenotypic data from both LR and DFS: 0.30 for bulls and 0.42 for cows. In fat yield, no improvement in R^2 was observed (0.18) for bulls and an unexpected decrease was observed for cows (0.21). The results showed that ssGBLUP was successfully implemented for the LR population with local and external data in the milk yield trait, but not in the fat yield trait.

The fourth objective was to implement the metafounder (MF) approach in ssGBLUP (Publ. III). The data were a subset of Finnish Red dairy cattle, including 112,479 cows with first lactation 305d milk records. Genomic data were obtained from 3,571 bulls and 16,186 cows. The 236 unknown parent groups assigned by the Nordic Cattle Genetic Evaluation (NAV) were reduced to 8 MFs. MF covariance matrix ($\boldsymbol{\Gamma}$) was created using base population allele frequencies estimated using a one-generation pedigree for each animal. After the estimation, markers were filtered with a minor allele frequency criterion of 0.05. Diagonal elements of the genomic relationship matrix had a lower correlation with the regular pedigree relationship matrix ($\mathbf{A}$; 0.66) than with the one using $\boldsymbol{\Gamma}$ ($\mathbf{A}^{\boldsymbol{\Gamma}}$; 0.76). Validation reliability of milk GEBVs in bulls increased by 0.04 (from 0.27 to 0.31) when using the MF approach. In cows the gain was 0.01 (from 0.36 to 0.37). The correlation of bull GEBVs between 8 MF and 236 UPG models was 0.972.

This thesis presents information needed to estimate ssGBLUP predictions in dairy populations with a low number of genotyped animals. Results can be used by LR farmers to improve data recording and implement genomic prediction. Furthermore, research on MFs should be used to improve the approach in ssGBLUP prediction.

Acknowledgements

This PhD thesis was the result of my long and indirect journey from veterinary science to animal breeding across countries and institutions. Thanks to many talented and kind-hearted people who made this journey possible.

I am grateful to institutions and authorities whose funding and support made this work possible, in particular Natural Resources Institute Finland (LUKE), Russian Research Institute of Farm Animal Genetics and Breeding (RRIFAGB), the University of Helsinki Department of Agricultural science, the doctoral program in Sustainable Use of Renewable Natural Resources, Committee for the Agro-Industrial and Fishery complex of Leningrad region, VikingGenetics, NAV, LLC Laboratoria Genome, and LLC Plinor.

I would like to acknowledge all my supervisors: Prof. Ismo Strandén, Prof. Esa Mäntysaari, and Prof. Pekka Uimari. Ismo, you are the most intelligent and humble person I know. Thank you for your wise advice, endless revision work, and patience in explaining practically everything to me. Esa, I certainly want to express my eternal gratitude for believing in a Russian student, countless support, and lessons on how to see the good even in bad situations and results. You always made me feel better and valuable. My main university supervisor Pekka, thank you for valuable advices, comments, and inputs to my research work. You made my university life and study curve easy and smooth.

I thank my co-supervisor Prof. Kirill Plemyashov for support and incredible help to establish scientific contacts all around the world. I am grateful for the constructive comments and opinions provided by Prof. Raphael Mrode and Prof. Mario Calus. Thank you, Dr. Ole Christensen, for agreeing to act as an opponent and Prof. Pekka Uimari for being my custos.

Adjunct Prof. Jarmo Juga, thank you for hosting me at the University and teaching me since the time I was a visitor. I am grateful for your job in the establishment of the Russian-Finnish scientific collaboration.

I would like to express my gratitude to my group manager Nina Schulman, for help and support during my research period at LUKE. A special thanks to my colleague Maria Leino and ex-colleague Timo Knurr for warm welcome and help in the early days at LUKE. I appreciate Maria's openness and daily non-scientific small talk. Thanks Timo Pitkänen and Hafedh Ben Zaabza for co-

authorship, joint work, endless discussions, and spare-time activities. I would like to acknowledge all my LUKE colleagues and especially Hanni Kärkkäinen, Riitta Kempe, Terhi Mehtiö, Minna Koivula, Kaarina Matilainen, Matti Taskinen, Martin Lidauer, Antti Kause, Enyew Negussie, and Marja-Liisa Sévon-Aimonen—thank you all for making me feel at home in Finland and at LUKE. I am very proud to be a part of such a powerful, united, and cutting-edge team.

Thanks to my former colleagues from RRIFAGB, especially Anna Petrova, Natalia Dementeva, and Elena Nikitkina. I would like to express gratitude to Ekaterina Saksa, who introduced me to the world of animal breeding, and my first group leader Mikhail Smaragdov, who showed me the world of molecular genetics. I sincerely thank my teachers and colleagues from University of Helsinki, especially Asko Mäki-Tanila, Kari Elo, and Katja Martikainen. Thanks to my many colleagues from NAV, but especially Jukka Pösö and Gert Aamand, for the co-authorship and experience and the knowledge they shared with me. Special thanks to Søren Borchersen from VikingGenetics for collaboration and support in establishing international cooperation.

Last but not least, I would like to thank my family for their faith in me and encouragement. There's is still time for a glass of wine and jazz record, dad. My deepest and warmest gratitude goes to my wife Anna, who was keeping our family nest all these long years. Your patience, faith, and unbelievable support made all this possible. Every time in endless trips, I knew a loving heart was waiting for me; thank you.

Content

List of original publications

This thesis is based on the following publications:

I Kudinov, A. A., Juga, J., Mäntysaari, E. A., Strandén, I., Saksa, E. I., Smaragdov M. G. and Uimari, P. 2018. Developing a genetic evaluation system for milk traits in Russian black and white dairy cattle. Agricultural and Food Science, 27: 85-95. https://doi.org/10.23986/afsci.69772

II Kudinov, A. A., Mäntysaari, E. A., Pitkänen T. J., Saksa E. I., Aamand G. P., Uimari, P. and Strandén, I. 2021. Single-step genomic evaluation of Russian dairy cattle using internal and external information. Journal of Animal Breeding and Genetics (submitted on 20.04.2021).

III Kudinov, A. A., Mäntysaari, E. A., Aamand G. P., Uimari, P. and Strandén, I. 2020. Metafounder approach for single-step genomic evaluations of Red Dairy cattle. Journal of Dairy Science, 103(7): 6299-6310. https://doi.org/10.3168/jds.2019-17483

The publications are referred to in the text by their roman numerals.

The author participated in research planning, data editing, statistical analyses, interpretation of results, and was the major writer of the above papers.

Abbreviations

AF	Allele frequencies
AI	Artificial insemination
AM BLUP	Animal model best liner unbiased prediction
CC	Contemporary comparison
CIS	The Commonwealth of Independent States
DFS	Denmark, Finland, and Sweden
DRP	Deregressed daughter performances
DYD	Daughter yield deviations
EBV	Estimated breeding value
EDC	Effective daughter contributions
ERC	Effective record contribution
FLM	First lactation model
FRDC	Finnish Red Dairy cattle
GEBV	Genomic enhanced breeding values
GLS	Generalized least squares
HOL	Holstein cattle
ICAR	The International Committee for Animal Recording
LR	Leningrad region of Russia
MACE	Multi-trait across-country evaluations
MAF	Minor allele frequencies
MF	Metafounder
MLM	Multiple lactation model
MME	Mixed model equation
NAV	Nordic evaluation service
RBW	Russian Black and White cattle
REML	Restricted maximum likelihood
RRIFAGB	Russian Research Institute of Farm Animal Genetics
SNP	Single-nucleotide polymorphism
ssGBLUP	Single-step genomic best linear unbiased prediction
UPG	Unknown parent groups
YD	Yield deviation
305d	305-day cumulative yield

1. Introduction

1.1 Genomic prediction in small populations

Over the past decade, genomic information has been successfully used to predict breeding values in dairy cattle (VanRaden, 2020), owing to which the dairy industry has notably changed. The generation interval has reduced by approximately 2.6 years, the fraction of genomically tested young candidate bulls at artificial insemination (AI) stations has reached 70% (Mäntysaari et al, 2020), and various farmers can obtain genetically outstanding bull candidates.

Genomic prediction is beneficial and attractive for all cattle breeds, but several aspects should be considered before its practical implementation. Precise phenotypic recording is the basis of accurate genomic prediction. In fact, genomic data does not replace but enhances the so-called traditional evaluation. The pedigree can be verified using genomic data, but it cannot be used to correct erroneous phenotypic records.

The number of genotyped animals with reliable estimated breeding values (EBVs) directly affect the accuracy of genomic prediction in candidate animals (Goddard et al., 2009). The original genomic evaluation approach predicts breeding values of the candidate animals using information derived from the genotyped reference population (Meuwissen et al., 2001). Progeny-tested bulls and elite cows have been the top choice for the reference population in large commercial breeds. Massive genotyping of young animals and collection of phenotypic records have notably improved prediction accuracy (Wiggans et al., 2017).

Genomic prediction is always challenging in small dairy breeds and populations because only a limited number of progeny-tested bulls are available. The first promising approach to achieving an adequate prediction accuracy is by including cows in the reference population (Ding et al., 2013; Li et al., 2014). Reliability of cow EBVs is always lower than that of progeny-tested bull EBVs because the number of descendants with records per cow is lower. Thus, to achieve the same prediction accuracy, a reference population including majorly cows needs to be larger than that including majorly bulls.

Therefore, more cows need to be genotyped. Massive genotyping of cows is an expensive approach to implement genomic prediction.

Another approach to increase the reference population size is to include data from an external related population. This will divide the costs and increase benefits (Lund et al., 2011; Jorjani et al., 2012; Ma et al., 2014). The collaboration may be based on only genomic or genomic and phenotypic data exchange. The most illustrative example of ongoing joint genetic and genomic evaluation is the Nordic cooperation (NAV, i.e., Denmark, Finland, and Sweden; https://www.nordicebv.info/). If sharing recorded data is undesirable, an alternative approach is to exchange EBVs through the multiple-trait, across-country evaluations (MACE Interbull, Uppsala, Sweden). Several methods to include external EBVs with corresponding reliability values into internal evaluations have been developed (VanRaden, 2001; 2012; Přibyl et al., 2013). Vandenplas et al. (2014) described an unified approach for combining internal data and pedigree information with external EBVs. The method used combined information, was free of double counting, and avoided overestimation of reliability. The method was successfully applied to genomic prediction models (Vandenplas et al., 2016, II).

1.2 Single-step genomic evaluations

Most dairy cattle evaluations use a multi-step approach to perform routine genomic predictions (Mäntysaari et al, 2020). The term multi-step refers to the need to perform several steps to predict genomic enhanced breeding values (GEBVs). In the first step, the approach estimates marker effect solutions using pseudo-observations and genotypes of the reference animals; in the second step, it predicts GEBVs of the candidate animals (VanRaden, 2008).

The single-step genomic best linear unbiased prediction (ssGBLUP) approach simultaneously uses genomic, pedigree, and phenotypic information to predict (G)EBVs of genotyped and non-genotyped animals (Aguilar et al., 2010; Christensen & Lund, 2010). The ssGBLUP approach is more elegant than the multi-step approach because ssGBLUP; in simple terms, represents the traditional best linear unbiased prediction animal model (BLUP-AM)

approach upgraded by genomic information. The ssGBLUP method predicts GEBVs more accurately than the multi-step approach when the population has a small fraction of genotyped animals (Christensen et al., 2012; Song et al., 2018). The advantage of the single-step approach over the multi-step approach has also been demonstrated in evaluations where external information was integrated (Přibyl et al., 2013). Incorporating external information in the multi-step approach would require an extra step, causing bias in GEBV prediction (Guarini et al., 2019)

Two theoretical assumptions impending ssGBLUP to be regarded as carefree: the same scale and equal base population of pedigree ($\mathbf{A}$) and genomic ($\mathbf{G}$) relationship matrices (Christensen et al., 2012). Several methods have been proposed to make $\mathbf{G}$ similar to $\mathbf{A}$, for example, the use of base population allele frequencies (AFs) (VanRaden, 2008) and scale and center of elements of $\mathbf{G}$ to have, on average, the same diagonal and off-diagonal elements as in $\mathbf{A}$ (Vitezica et al., 2011; Christensen et al., 2012). In practice, base population AF are unknown, and the $\mathbf{G}$ matrix is often constructed using AF observed in the genotyped population. Dairy cattle pedigree can seldom be traced to a genetically homogeneous base population because the pedigree often has a complicated breed structure with unknown parent information (VanRaden, 1992; Sponenberg & Bixby, 2007). However, the completeness of the pedigree is critical to the consistency between $\mathbf{G}$ and $\mathbf{A}_{22}$ (sub-block of $\mathbf{A}$) matrix (Misztal et al., 2010; 2013).

The metafounder (MF) approach was proposed by Legarra et al. (2015) to achieve compatibility in the pedigree and genomic relationship matrices. The MF approach combines Christensen's (2012) idea of using AF equal to 0.5 for all markers in the $\mathbf{G}$ matrix and assigning unknown parents to pseudo-individuals with self-relationships in the $\mathbf{A}$ matrix. The MFs are related base populations with non-zero inbreeding coefficients. The relationships within and between the MFs are presented by gamma matrix ($\mathbf{\Gamma}$). The $\mathbf{\Gamma}$ matrix is used to compute the $\mathbf{A}^{\mathbf{\Gamma}}$ from $\mathbf{A}$ matrix. The $\mathbf{\Gamma}$ matrix may be constructed using an estimated base or observed genotyped population AF (e.g., Legarra et al., 2015; Garcia-Bacciano et al., 2017). The method has provided promising results in

the analysis of multiple breed pig pedigree (Xiang et al., 2017) and simulated data (Garcia-Baccino et al., 2017). However, implementation of the MF approach can be challenging when the population has breeds with a high admixture. When the number of unknown parent groups (UPGs) is large, a UPG may be associated with a low number of rare allele genotypes. The $\boldsymbol{\Gamma}$ matrix may be poorly estimated when certain AFs are estimated inaccurately owing to the low number of rare alleles.

1.3 Dairy cattle breeding in the Leningrad region

The Leningrad region (LR) is in the Northwestern Federal District of Russia, located in the eastern bay of the Baltic Sea, bordering Finland and Estonia. For many years, the LR has been the leading dairy region in Russia according to average milk production per cow (8,681 kg in 2017, Yearbook 2017), with a high level of integration of modern technologies in the agricultural sector.

The most popular dairy breeds of the region are Russian Black and White (RBW) and Holstein (HOL). The RBW breed was created by crossbreeding native Russian breeds with imported Dutch bulls in various parts of Russia since the 1820s. In 1925, the crosses were improved by massive importation of Ost-Frisian bulls from Germany, Estonia, Lithuania, and the Netherlands. Finally, in 1959, RBW was registered as a breed, and pure breeding started (Arzumanyan, 1973). In the last four decades, farmer preference for dual-purpose breeds (dairy–beef) was changed to dairy-purpose breeds. Breed improvement has been done through importation of HOL semen, young calves, and heifers. Currently, it is challenging to distinguish between the LR RBW and HOL breeds as they have high admixture (Smaragdov, 2018).

Approximately 50 breeding herds and reproducers in LR are keeping LR RBW and HOL cattle. The herds are large, containing 800–4000 animals. The milk records are collected by technicians monthly and transmitted to the regional data center Plinor LLC (https://plinor.spb.ru/). None of the Russian data-processing centers are fully approved by ICAR (The International Committee for Animal Recording; https://www.icar.org/index.php/about-us-icar-facts/icar-members/, 2020). However, farms use milking robots and

equipment certified by ICAR, and milk laboratories analyse milk samples on the certified measuring devices.

The current official evaluation method used for breeding value estimation is Contemporary Comparison (CC; Instruction, 1979). The method was discontinued by many countries in the 1980s (Schaeffer, 2013) as it allows breeding value estimation only for bulls with daughters and does not work properly in herds with different environmental conditions. Several attempts have been made to apply the BLUP methods for bull and cow evaluation in LR (Shkirando, 1986; Ignashkina & Kuznetsov, 1988; Myakoshina et al., 1992). However, none of the BLUP-based methods were implemented at the industry level.

Young calves from the LR breeding herds are actively bought by various AI stations situated in Russia and the Commonwealth of Independent States countries. Large AI stations nowadays ask farmers to provide GEBVs of bull candidates before a bargain. Update of the evaluation system from CC to BLUP and further to ssGBLUP would be beneficial for farmers and breeding companies. Transmission of the genetics from the best dairy regions to the regions with abundant feed and land resources would increase the level of milk production in Russia. To introduce a modern genomic evaluation system, in 2015, the Leningrad Committee on Agriculture and Fishery started a 3-year research and development project on ssGBLUP evaluations. As a consequence, a project (RUGE) was established between the Russian Research Institute of Farm Animal Genetics (RRIFAGB) and Breeding, University of Helsinki, and Natural Recourses Institute Finland.

2. Objectives

The main objective of this thesis was to find ways to implement and improve ssGBLUP in a small-scale dairy cattle population. The specific aims were to predict GEBVs for the LR dairy cattle using local genotypes, enhance the reliability of prediction by incorporating data from an external breeding population, and implement the MF approach in an ssGBLUP dairy cattle evaluation.

The steps to achieve these objectives were as follows:

1) to prepare LR phenotypic and pedigree data for genetic and genomic prediction (I);
2) to estimate variance components in BLUP-AM for milk, fat, and protein yield in RBW and HOL cattle of the LR (I, II);
3) to test ssGBLUP using genotypes of LR bulls and cows (II);
4) to enhance the reliability of genomic prediction by using external Nordic (Denmark, Finland, and Sweden) HOL genotypes and MACE EBVs (II);
5) to test the MF approach in ssGBLUP using Finnish Red dairy cattle (FRDC) data (III);
6) to develop an optimal $\boldsymbol{\Gamma}$ matrix for FRDC pedigree (III).

3. Materials and methods

3.1 Data sets

Three data sets were used in the studies included in the thesis (Table 1). 1) LR phenotypic (I, II) and genomic (II) data from RBW and HOL breeding herds; 2) MACE EBVs and genotypes of Nordic (DFS) HOL bulls (II); and 3) FRDC phenotypic and genomic data (III).

Table 1. Distribution of the data by original publication

Data set[1]	№ records	Birth years	№ genotypes	Publication
LR	320,633	2000–2013	-	I
LR	363,833	2000–2015	1,507	II
LR gDFS	363,833	2000–2015	1,507 + 414	
LR DFS	363,833 + 487	1960–2015	1,507 + 414	
FRDC	112,479	1988–2018	19,757	III

[1]LR - Leningrad region;
gDFS[1] - includes genotypes from Nordic bulls only;
DFS[2] - includes genotypes and estimated breeding values from Nordic bulls only;
FRDC – Finnish Red dairy cattle

3.1.1. Leningrad region data

Phenotypic data of HOL and RBW breeds were obtained from regional recording centre LLC Plinor. The data were collected for 49 herds for cows born during the period 2000–2013. Records contained information about 305d milk, fat, and protein yields (kg) and dates of birth, service, and calving from all available lactations. The data were edited to exclude missing values, unfinished lactations, outliers, and cows with missing first lactation (I). The highest parity allowed in the variance component estimation and the EBV prediction was 3 and 5, respectively. The final data set included 320,633 records (Table 2; I). Furthermore, the data were supplemented by 43,200 up-to-date records (Table 2; II). The pedigree included 1,779 sires, 159,069 cows, and 59,774 dams without their own records (II).

Table 2. Number of milk records by lactation in LR data.

Lactation	Number of milk records	
	Publication I	Publication II
1	141,868	158,838
2	91,269	102,836
3	51,239	57,835
4	25,298	28,684
5	10,959	15,640
Total	320,633	363,833

Genomic data were raw single nucleotide polymorphism (SNP) marker data from 1,080 cows and 427 bulls obtained from repositories of RRIFAGB and LLC Laboratoria Genome. Genotyped cows originated from 13 herds, with the average (± standard deviation; SD) number of cows per herd 82 ± 21. The following criteria were used to perform SNP quality control: call rate >95% and minor allele frequency >5%. After imputation, 43,194 markers remained for further genomic prediction (II).

Two reduced data sets were created for the calculation of validation reliability and bias of genomic prediction (II). For the bull validation test, milk, and fat records from the last four production years (2012–2015) were removed. An exception was made for the genotyped cows not closely related (i.e., not daughters, granddaughters, or sibs) to the validation bulls and representing contemporary groups (herd–year–season) with at least five animals. The data records from these cows were retained in order to avoid exhaustion of the training set. The bull validation test set had 48 bulls with effective daughter contribution (EDC) greater than 20 in the full data set but zero EDC in the reduced data set. For the cow validation test, records from the last production year (2015) were excluded. There were 221 test cows that had no records in the reduced data set but at least one record in the full data set.

3.1.2. Nordic data

For 487 bulls present both in the LR and DFS evaluations, the MACE EBVs on the DFS scale were obtained. Only bulls with more than 20 daughters were used. The EDC values were computed from MACE EBV reliabilities using the

reverse reliability estimation (Taskinen et al., 2014) and the LR heritability values (II). Using the calculated EDC and full pedigree information, the MACE EBVs were converted to deregressed daughter performance (DRP) values (Jairath et al., 1998; Strandén & Mäntysaari, 2010).

Genotypes were available for 414 of the 487 bulls. Genotype quality control and imputation were done by the NAV. The number of markers used was the same as that in the LR data (43,194).

3.1.3. Finnish Red dairy cattle data

The FRDC data were extracted from the DFS production evaluation database and treated as a small-scale population. The Finnish herds included in the study had at least 10 genotyped cows with records. There were 112,479 first-lactation 305d milk yield records from 1988–2018 for 426 herds. The pedigree included 226,012 animals born in 1960–2016 and 236 UPGs assigned by NAV. The groups were based on selection path, birth year, and population of origin (III).

Genomic data were presented by 46,914 markers from 3,571 bulls and 16,186 cows (III).

Cow and bull validation data sets were created by excluding the milk yield records for either the last year (2018) or for the last four production years (2015-2018), respectively. The bull validation test set had 101 bulls with EDC greater than 20 in the full data set but zero EDC in the reduced data set. The cow validation test set had 3,551 cows which had no records in the reduced data set but one record in the full data set.

3.2 Statistical models

3.2.1 Leningrad region BLUP-AM

Single-trait BLUP-AM with either first lactation records (first lactation model; FLM) or multiple lactation records (multiple lactation model; MLM) were used to estimate milk, fat, and protein variance components and EBVs (I).

The repeatability model for milk, fat, and protein was as follows:

$$\boldsymbol{y_{ijk} = DOAC_i + HYS_j + a_k + p_k + e_{ijk},}$$

where $\boldsymbol{y_{ijk}}$ is the yield observation of $\boldsymbol{ijk}^{\text{th}}$ cow; $\boldsymbol{DOAC_i}$ is the $\boldsymbol{i}^{\text{th}}$ fixed effect, days open by age of calving by lactation; $\boldsymbol{HYS_j}$ is the $\boldsymbol{j}^{\text{th}}$ fixed effect, herd by year by season; $\boldsymbol{a_k} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{A}\sigma_a^2)$ is the random additive genetic value of the $\boldsymbol{k}^{\text{th}}$ animal; $\boldsymbol{p_k} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I}\sigma_p^2)$ is the random permanent environment effect associated with the $\boldsymbol{k}^{\text{th}}$ animal; and $\boldsymbol{e_{ijk}} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$ is the residual effect. **A** and **I** are relationship and incidence matrices, and σ_a^2, σ_p^2, and σ_e^2 correspond to additive genetic, permanent environment, and residual variances, respectively. The permanent environment effect ($\boldsymbol{p}$) was not included in FLM.

The DOAC effect was obtained by combining the days open (DO) and the age of calving (AC) classes within lactation. The HYS effect was obtained as a combination of herd code, year, and season (coded as 1 to 4). The number of levels in the DOAC and the HYS effects were 203 and 2603, respectively (I).

Unknown parent groups were added to the pedigree and treated as a random effect. Six groups were created on the basis of selection path and place of origin with subsequent subgrouping by year. The final number of groups was 218 (I).

3.2.2 Leningrad region ssGBLUP models

Use of the best bulls in a few top herds can lead to an interaction between sire and herd (Dimov et al., 1995). The single-trait mixed model equation (MME; I) was modified by including the random effect herd by sire interaction and used to estimate milk and fat yield variance components and (G)EBVs (II). The model notation was changed to the following:

$$\boldsymbol{y_{imjlk} = DOAC_i + HYS_{mj} + hs_{ml} + a_k + p_k + e_{imjlk}}$$

where $\boldsymbol{y_{imjlk}}$ is the yield observation of cow $\boldsymbol{k}$; $\boldsymbol{DOAC_i}$ is the $\boldsymbol{i}^{\text{th}}$ fixed effect, days open by age of calving; $\boldsymbol{HYS_{mj}}$ is the fixed effect, herd $\boldsymbol{m}$ in the year by season $\boldsymbol{j}$; $\boldsymbol{hs_{ml}} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I}\sigma_{hs}^2)$ is the random herd $\boldsymbol{m}$ by sire $\boldsymbol{l}$ interaction effect; $\boldsymbol{a_k} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{A}\sigma_a^2)$ is the random additive genetic value of the $\boldsymbol{k}^{\text{th}}$ animal; $\boldsymbol{p_k} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I}\sigma_p^2)$ is the random permanent environment effect associated with the

k^{th} animal; and $\boldsymbol{e_{imjlk}} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$ is the residual effect. $\boldsymbol{A}$ and $\boldsymbol{I}$ are relationship and incidence matrices, and σ_{hs}^2, σ_a^2, σ_p^2, and σ_e^2 correspond to herd by sire interaction and additive genetic, permanent environment, and residual variances, respectively.

Number of UPGs was reduced to 54 by combination of time intervals into larger classes and revision of selection path (II).

In an ssGBLUP model with UPG, the joint inverse relationship matrix of genotyped and non-genotyped animals was as follows:

$$\mathbf{H^{-1}} = \mathbf{A}_{\text{UPG}}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & -\mathbf{BQ_2} \\ \mathbf{0} & -\mathbf{Q_2'B} & \mathbf{Q_2'BQ_2} \end{pmatrix},$$

where $\mathbf{B} = \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}}$, $\mathbf{G}$ is the genomic relationship matrix, $\mathbf{A_{22}}$ is a subset of the pedigree relationship matrix ($\mathbf{A}$) including genotyped animals only;

$$\mathbf{A_{UPG}^{-1}} = \begin{pmatrix} \mathbf{A^{11}} & \mathbf{A^{12}} & -(\mathbf{A^{11}Q_1} + \mathbf{A^{12}Q_2}) \\ \mathbf{A^{21}} & \mathbf{A^{22}} & -(\mathbf{A^{21}Q_1} + \mathbf{A^{22}Q_2}) \\ -(\mathbf{Q_1'A^{11}} + \mathbf{Q_2'A^{21}}) & -(\mathbf{Q_1'A^{12}} + \mathbf{Q_2'A^{22}}) & \mathbf{Q'A^{-1}Q} \end{pmatrix};$$

$\mathbf{Q}$ represents proportions of contributions each animal receives from the UPG; $\mathbf{Q_1}$ and $\mathbf{Q_2}$ are submatrices of $\mathbf{Q}$ corresponding to the non-genotyped and genotyped animals, respectively; and $\mathbf{A}^{ij}$ is submatrix of $\mathbf{A^{-1}}$ with superscript (i or j) value 1 for non-genotyped and value 2 for genotyped animals. Inbreeding coefficients were used in the calculations of the inverse pedigree-based relationship matrices $\mathbf{A^{-1}}$ and $\mathbf{A_{22}^{-1}}$.

It was assumed that the genotypes model 90% of the genetic variance. The genomic relationships were regressed towards the pedigree relationships using the following equation:

$$\mathbf{G} = s_t\,(1-w)\,\mathbf{G}_{05} + w\mathbf{A}_{22},$$

where w signifies the residual polygenic effect equal 0.1, and $\mathbf{G_{05}} = 2\left(\frac{\mathbf{M}_{101}\mathbf{M}_{101}'}{\text{m}}\right)$, with $\mathbf{M}_{101}$ as an n by m marker matrix with the genotypes coded by {-1,0,1}, where m is the number of SNP markers and n is the number of

genotyped animals. The scaling factor $s_t = (\frac{trace(\mathbf{A_{22}})}{trace(\mathbf{G_{05}})})$, was used to assure that the diagonals of the **G** matrix remained on average the same as the diagonals of the $\mathbf{A}_{22}$ matrix. The genomic relationship matrix was not computed using AF based on observed genotypes because the low number of genotyped animals sparsely and non-uniformly distributed across years could lead to biased and erroneously defined AF base. Instead AF equal to 0.5 was used which suggests a distant base population and less biased genomic relationships among animals.

For notation simplicity, the MME of integration of DFS information into the LR evaluation presented with only the additive genetic effect and the fixed effects was expressed as follows:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + H^{-1}}\,\sigma_a^{-2} + \mathbf{D_N^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y + D_N^{-1}a_N^*} \end{bmatrix},$$

where **X** is a design matrix relating the fixed effects (**DOAC** and **HYS**) to the records, **Z** is design matrix relating the random effects to the records, $\mathbf{R} = \mathbf{I}\sigma_e^2$ is the residual (co)variance matrix, **D** is the diagonal matrix with the EDC increase for bulls due to the DFS data and EDC zero for cows, subscript *N* pertains to the DFS MACE evaluation, **b** is a vector of the fixed effects, $\mathbf{a} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{A}\sigma_a^2)$ is a vector of random animal genetic effect, $\mathbf{a_N^*}$ is a vector of the DRP from the DFS MACE evaluation, and $\boldsymbol{y}$ is a vector of milk or fat yield records (II).

To perform validation of the model fit, the full data were used to calculate daughter yield deviations (DYD) for bulls and yield deviations (YD) for cows using the corresponding ssGBLUP model. Bias was estimated by (G)EBV overdispersion, that is, the regression coefficient b_1 in the validation regression model (D)YD $= b_0 + b_1$ GEBV, and by average difference between GEBV and (D)YD. The DYD observation for bull $\boldsymbol{i}$ were weighted using $\boldsymbol{k_i}$ calculated as $\boldsymbol{k_i} = \frac{EDC_i}{EDC_i + \lambda_1}$, where $\lambda_1 = (4 - h^2)/h^2$. The YD observations for cow $\boldsymbol{j}$ were weighted using parameter k_j calculated as $k_j = \frac{ERC_j}{ERC_j + \lambda_2}$, where $\lambda_2 = (1 - h^2)/h^2$, and ERC_j is the effective record contribution of cow j calculated as

$ERC_j = \frac{1-h^2}{h^2} \times \frac{r_j^2}{1-r_j^2}$, where r_j^2 = reliability of cow j EBV (Přibyl et al., 2013). The within herd heritability was calculated using the formula $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_{pe}^2 + \sigma_e^2)$, where σ_a^2, σ_{pe}^2, and σ_e^2 are genetic, permanent environmental, and residual variances, respectively. The validation reliability (R^2) was calculated as the squared correlation between (D)YD and the reduced data set GEBVs divided by average k_i and k_j for bulls and cows, respectively.

The LR ssGBLUP evaluations were implemented and tested using three scenarios (II). In the first scenario (ssLR), the LR phenotypic and genomic data were used. In the second scenario (ssLRg), the ssLR was upgraded by DFS genotypes. In the third scenario (ssLRdfs), the ssLR was upgraded to include both DFS genotypes and DRP values. Thus, ssLRg included more genomic information than ssLR, whereas ssLRdfs included more phenotypic information than ssLRg.

3.2.3 Finnish Red Dairy cattle single-step GBLUP models

In an ssGBLUP model with UPG, the joint inverse relationship matrix of genotyped and non-genotyped animals was expressed as follows:

$$\mathbf{H^{-1}} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G_{PvR1}^{-1}} - \mathbf{A_{22}^{-1}} \end{pmatrix},$$

where $\mathbf{A}$ is the full pedigree relationship matrix; $\mathbf{G_{PvR1}}$ is the genomic relationship matrix constructed using VanRaden (2008) method 1, where base population AFs were used to centre and scale the marker data; and $\mathbf{A_{22}}$ is a pedigree relationship matrix of genotyped animals. Mean genetic levels of animals with missing parental information were modelled using pedigree based UPG. In the model, unknown parents were assumed to be unrelated and completely outbred. The base population AFs were estimated with the generalized least squares (GLS) model (McPeek et al., 2004). The genomic information was assumed to account for 90% of the variation in breeding values.

In an ssGBLUP model with the MF, the matrix $\mathbf{H}^{-1}$ was replaced by a modified $(\mathbf{H^{\Gamma}})^{-1}$computed as follows:

$$(\mathbf{H}^{\boldsymbol{\Gamma}})^{-1} = (\mathbf{A}^{\boldsymbol{\Gamma}})^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G_w^{-1}} - \left(\mathbf{A}_{\mathbf{22}}^{\boldsymbol{\Gamma}}\right)^{-1} \end{pmatrix},$$

where $\mathbf{G_w} = (1-w)\mathbf{G_{05}} + w\mathbf{A}_{\mathbf{22}}^{\boldsymbol{\Gamma}}$, w is the residual polygenic effect equal 0.1, $\mathbf{G_{05}} = (\mathbf{M_{101}}\mathbf{M}'_{\mathbf{101}})\frac{2}{\mathrm{m}}$, $\mathbf{M}_{101}$ is an n by m marker matrix with genotypes coded by {-1,0,1}, m is the number of SNP markers, n is the number of genotyped animals, $\mathbf{A}^{\boldsymbol{\Gamma}}$ is the pedigree relationship matrix formed with a $\boldsymbol{\Gamma}$ matrix, and $\mathbf{A}_{22}^{\boldsymbol{\Gamma}}$ is a submatrix of $\mathbf{A}^{\boldsymbol{\Gamma}}$ for genotyped animals. The variance–covariance structure of MFs was estimated by $\boldsymbol{\Gamma} = \mathbf{8}\,\mathbf{Cov}(\mathbf{P})$, where $\mathbf{P}$ is an m by r matrix of base population AFs and r is the number of MFs.

The number of UPGs was downscaled from 236 to 8. The eight groups were as follows: six FRDC groups (birth years <1971, 1971–1980, 1981–1990, 1991–2000, 2001–2010, and 2011–2016), one HOL, and one OTHER breed groups. In the MF approach, the eight UPGs were treated as MFs. AFs were estimated for each MF using the GLS model: $\boldsymbol{m}_i = \mathbf{U}\boldsymbol{\mu}_i + \boldsymbol{e}_i$, where $\boldsymbol{m}_i$ is an n by 1 vector of marker i genotypes; $\mathbf{U}$ is an n by 8 matrix, the rows of which sum to 1 and which assigns genotyped individuals to fractions of MF; $\boldsymbol{\mu_i}$ is an 8 by 1 vector of group means; and $\boldsymbol{e}_i \sim (\mathbf{0}, \mathbf{A}_{22}^{*}\sigma^2)$, where $\mathbf{A}_{22}^{*}$ is the pedigree relationship matrix for the genotyped animals and σ^2 is the common variance. In allele frequency estimation, the common variance need not be known. Estimated base population AFs for the MF are $\widehat{\boldsymbol{p}}_i = \frac{1}{2}\widehat{\boldsymbol{\mu}}_i$ for each marker $i = 1, \ldots, m$. Estimated eight columns of AFs in $\widehat{\boldsymbol{p}}_i$ were used to calculate the $\boldsymbol{\Gamma}$ matrix. The variance of breeding values in base populations descending from MFs were calculated using the correction factor k, that is, $\sigma_{a,k}^2 = \sigma_a^2/k$, where $k = (1 + \boldsymbol{tr}(\boldsymbol{\Gamma})/(2n) - 1'\boldsymbol{\Gamma}1/n^2)$, and tr($\boldsymbol{\Gamma}$) is the sum of diagonal elements of the $\boldsymbol{\Gamma}$ matrix.

To estimate the AFs for the MFs in the GLS model, the $\mathbf{A}_{\mathbf{22}}^{*}$ matrix was based on a truncated pedigree, where one parent generation at most was defined for the genotyped animals.

The effect of minor allele frequencies (MAFs) on the MF covariances were tested by creating two $\boldsymbol{\Gamma}$ matrices. In the first matrix, the full $\mathbf{P}$ matrix was used to calculate the $\boldsymbol{\Gamma}$ matrix, denoted $\boldsymbol{\Gamma}_8$. In the second matrix, denoted $\boldsymbol{\Gamma}_{8\mathrm{MAF}}$,

only those markers with MAF ≥0.05 in all FRDC cattle MF were included in the **P** matrix. The MAF requirement eliminated 3,783 markers and left 43,131 markers that were used to calculate the $\boldsymbol{\Gamma}_{8\mathrm{MAF}}$ matrix. Consequently, two correction factors ($k_{\boldsymbol{\Gamma}_8}$ and $k_{\boldsymbol{\Gamma}_{8\mathrm{MAF}}}$) were calculated using $\boldsymbol{\Gamma}_8$ and $\boldsymbol{\Gamma}_{8\mathrm{MAF}}$, respectively.

To perform model validation, DYD and YD were calculated using full data and animal model. Regression models for bulls and cows were $(\mathrm{D})\mathrm{YD} = \mathrm{b}_0 + \mathrm{b}_1 * \mathrm{GEBV}$, with weights for the DYD observations. The weight for DYD was $\mathrm{EDC}/(\mathrm{EDC} + \lambda)$, where λ is $(4 - \mathrm{h}^2)/\mathrm{h}^2$ and, h^2 is the heritability equal to 0.44. To reach adjusted validation reliability, we divided the model coefficient of determination (R^2) by the average weight. The regression coefficient b_1 for the bulls was multiplied by two because DYD only represents half of the breeding value of the sire.

Four ssGBLUP models were tested: in the UPG approach, the models were with original 236 ($\mathrm{ssGBLUP}_{236\mathrm{UPG}}$) and newly defined 8 ($\mathrm{ssGBLUP}_{8\mathrm{UPG}}$) UPGs; in the MF approach, models were with $\boldsymbol{\Gamma}_8$ ($\mathrm{ssGBLUP}_{\boldsymbol{\Gamma}_8}$) and $\boldsymbol{\Gamma}_{8\mathrm{MAF}}$ ($\mathrm{ssGBLUP}_{\boldsymbol{\Gamma}_{8\mathrm{MAF}}}$).

3.3 Software

Pedigree pruning, calculation of inbreeding coefficients, and relationship submatrices $\mathbf{A_{22}}$ and $\mathbf{A}_{22}^{*}$ were performed using RelaX2 program (Strandén & Vuori, 2006). Variance components were estimated by restricted maximum likelihood (REML, Patterson & Thompson, 1971) in DMU software (Madsen et al., 2010) using AI-REML algorithm. Imputation of missing alleles was done using FImpute v. 2.2 software (Sargolzaei et al., 2014).

The $\mathbf{G^{-1}}$ and $\mathbf{B}$ matrices were computed using HGinv v. 0.87 program (Strandén & Mäntysaari, 2018). Base population AFs were estimated with the GLS model (McPeek et al., 2004) using the Bpop v. 0.30 program (Strandén & Mäntysaari, 2020). The EBVs and GEBVs and DRP, (D)YD, EDC, and ERC values were computed using MiX99 software (Strandén & Lidauer, 1999).

4. Main results

4.1 Leningrad region BLUP-AM

4.1.1 Phenotypic data and variance components

In the final data set, the average milk, fat, and protein yield was 7,644–8,165 kg, 296–317 kg, and 252–271 kg, respectively, depending on lactation number. The lowest average yield for all three traits was observed in the first lactation. The highest average yield was obtained in the second lactation for milk and fat and in the third lactation for protein.

The heritability estimates for milk, fat, and protein obtained using the FLM (0.24 ± 0.008,0.20 ± 0.007, and 0.20 ± 0.008) were higher than those obtained using the multiple lactation model (0.18 ± 0.004, 0.17 ± 0.005, and 0.19 ± 0.004). The repeatability estimates were 0.34 for milk and 0.31 for fat and protein (Table 3).

Table 3. Estimates of variance components for milk, fat, and protein yield traits

Model	Trait	σ_a^2	σ_p^2	σ_e^2	h^2	r
FLM	Milk	301,465	-	936,884	0.24	-
	Fat	328	-	1,353	0.20	-
	Protein	223	-	915	0.20	-
MLM	Milk	313,916	281,706	1,138,386	0.18	0.34
	Fat	412	324	1.630	0.17	0.31
	Protein	295	210	1.074	0.19	0.31

σ_a^2 – genetic variance, σ_p^2 - permanent environmental variance, and σ_e^2- residual variance components; h^2 - heritability; r – repeatability; FLM – first lactation model; MLM – multiple lactation model

4.1.2 Estimated breeding values

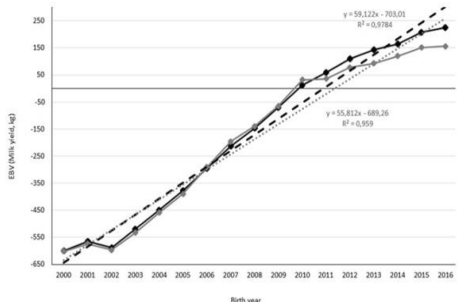
The average milk yield EBV calculated using FLM and MLM for cows born between 2000 and 2016 was 59 and 56 kg/year, respectively (Figure 1a). Both genetic trends were quite similar until 2010; however, MLM predicted a slower genetic trend after 2011 than FLM. The difference in the trend between the two

models was 3 kg, which is smaller than the Interbull criterion of $0.02 * \sigma_a^2$ (11.2 kg).

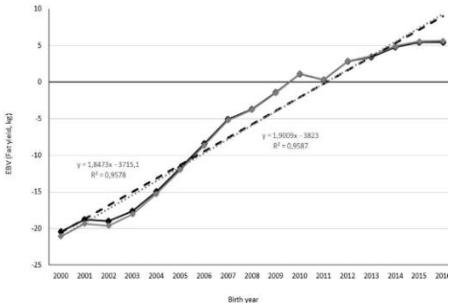
The average genetic gain estimated by FLM and MLM was nearly identical: 1.90 and 1.84 kg/year for fat yield (Figure 1b) and 1.67 and 1.62 kg/year for protein yield (Figure 1c), respectively. Both fat and protein yield trends plateaued from 2012 onward. The difference in the average genetic gain between FLM and MLM was 0.06 kg for fat and 0.05 kg for protein, which is smaller than the Interbull criterion of $0.02 * \sigma_a^2$ (0.41 and 0.34 kg, respectively)

Figure 1. Genetic trends for milk (a), fat (b), and protein (c) yield for cows and corresponding regression lines with equation showing annual response using a first lactation model (FLM; black lines) and multiple lactation model (MLM; gray lines).

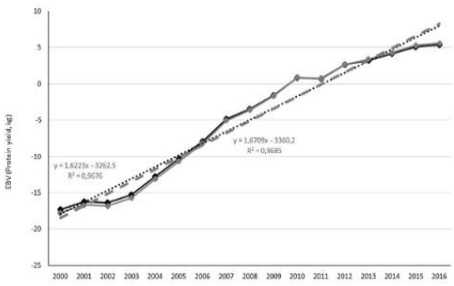
a)



b)



c)



4.2 Leningrad region ssGBLUP

4.2.1 Variance components and genomic estimated breeding values

Variance components estimated and used in the ssGBLUP runs are presented in Table 4. Heritability estimates for milk and fat yield were similar (0.21 ± 0.005). The repeatability estimate for milk yield was slightly higher (by 0.04) than that for fat yield.

Table 4. Estimates of variance components for milk and fat yield traits obtained from the mixed model equation with the random effect herd by sire interaction.

Trait	σ_a^2	σ_p^2	σ_{hs}^2	σ_e^2	h^2	r
Milk	330,735	274,195	80,532	955,257	0.21	0.39
Fat	451	300	118	1,393	0.21	0.35

σ_a^2 - genetic variance, σ_p^2 - permanent environmental variance, σ_{hs}^2 - herd-sire variance, and σ_e^2 - residual variance components; h^2 - heritability; r – repeatability

For bulls with EDC ≥20, the ssLR and ssLRg models showed the same average annual genetic change in milk yield (40 kg) in 1995–2010. However, the ssLRdfs model showed a higher average annual genetic change (60 kg) for the same period. A similar pattern was observed for fat yield; the estimated annual genetic change was 1.2 kg in ssLR and ssLRg and 1.9 kg in ssLRdfs. For cows, the average annual increase in milk yield was 50 kg in ssLR and ssLRg and 55 kg in ssLRdfs. For cows, the predicted change in fat yield was 1.7 kg in ssLR and ssLRg models and 1.9 kg in ssLRdfs.

4.2.2 Validation of the model fit

The highest validation reliability (R^2) for milk yield was observed in the ssLRdfs model—0.30 for bulls and 0.42 for cows (Table 5). However, the model had the lowest regression coefficient (b_1) for bulls and cows: 0.58 and 1.14, respectively. For cows, regression coefficients in all models were >1. In ssLR and ssLRg, the regression coefficients for bulls were close and <1 (0.78 and 0.80).

Table 5. Milk yield regression analysis results from the three single-step genomic best linear unbiased prediction (ssGBLUP) models in the Leningrad region (LR) Holstein and Russian Black and White cattle population

	Validation animals					
	Bulls (42 animals)			Cows (221 animals)		
Model	E (GEBV–DYD)	$2 * b_1$	R^2	E (GEBV–YD)	b_1	R^2
ssLR	529	0.78	0.21	65	1.69	0.38
ssLRg	557	0.80	0.21	91	1.55	0.36
ssLRdfs	748	0.58	0.30	113	1.14	0.42

GEBV – genomic enchanted breeding value; (D)YD – (daughter) yield deviation; E (GEBV–(D)YD) - average difference between GEBV and (D)YD; b_1- regression coefficient (multiplied by 2 in bulls because DYD represents half of the breeding value of the sire); R^2 - validation reliability; ssLR - ssGBLUP model with only LR data; ssLRg - ssGBLUP model with LR and DFS genomic data; ssLRdfs - ssGBLUP model with LR and DFS phenotypic and genomic data

In the fat yield for the bulls, the highest validation reliability was 0.18 reached by ssRLg and ssRLdfs models (Table 6). For the cows, the highest R^2 was archived by ssLR model. The addition of external genomic and phenotypic data reduced the validation reliability for the cows. Regression coefficient for bulls were lower than obtained from milk yield: 0.64, 0.68, and 0.41 in ssLR, ssLRg, and ssLRdfs, respectively. In cows, the b_1 coefficient was above one in in ssLR and ssLRg models: 1.86 and 1.67, respectively. In ssLRdfs, b_1 was below one (0.89).

Table 6. Fat yield regression analysis results from the three single-step genomic best linear unbiased prediction (ssGBLUP) models in the Leningrad region (LR) Holstein and Russian Black and White cattle population.

	Validation animals					
	Bulls (42 animals)			Cows (217 animals)		
Model	E (GEBV-DYD)	$2 * b_1$	R^2	E (GEBV-YD)	b_1	R^2
ssLR	18	0.64	0.17	6	1.86	0.41
ssLRg	19	0.68	0.18	7	1.67	0.34
ssLRdfs	27	0.41	0.18	7	0.89	0.21

GEBV – genomic enchanted breeding value; (D)YD – (daughter) yield deviation; E (GEBV -DYD) - average difference between GEBV and DYD; b_1 - regression coefficient (multiplied by 2 in bulls because DYD represents half of the breeding value of the sire); R^2- validation reliability; ssLR - ssGBLUP model with only LR data; ssLRg - ssGBLUP model with LR and DFS genomic data; ssLRdfs - ssGBLUP model with LR and DFS phenotypic and genomic data

4.3 Metafounder approach in single-step genomic evaluations

4.3.1 Gamma and relationship matrices

Elements of the $\boldsymbol{\Gamma}_{8MAF}$ matrix showed slightly smaller values than the corresponding elements in the $\boldsymbol{\Gamma}_8$ matrix owing to the inclusion of MAF threshold to select markers (Figure 2). All diagonal elements in the $\boldsymbol{\Gamma}$ matrices were <1, which correspond to negative inbreeding coefficients for the MFs. The average mean correlation between the FRDC and HOL MFs was 0.564 and 0.473 in $\boldsymbol{\Gamma}_8$ and $\boldsymbol{\Gamma}_{8MAF}$, respectively. The highest values of diagonal elements (self- correlation) were observed in the groups FRDC <1970 (0.618 and 0.719 in $\boldsymbol{\Gamma}_{8MAF}$ and $\boldsymbol{\Gamma}_8$, respectively) and OTHER (0.740 and 0.797 in $\boldsymbol{\Gamma}_{8MAF}$ and $\boldsymbol{\Gamma}_8$, respectively; III).

Figure 2. Estimated $\boldsymbol{\Gamma}_8$ (lower triangle) and $\boldsymbol{\Gamma}_{8MAF}$(upper triangle)

	FRDC <1970	FRDC 1971–1980	FRDC 1981–1990	FRDC 1991–2000	FRDC 2001–2010	FRDC 2011–2016	HOL	OTHER
FRDC <1970	0.618 (0.719)	0.555	0.563	0.563	0.566	0.566	0.471	0.453
FRDC 1971–1980	0.659	0.569 (0.670)	0.566	0.561	0.564	0.562	0.473	0.454
FRDC 1981–1990	0.668	0.670	0.609 (0.710)	0.588	0.589	0.585	0.473	0.452
FRDC 1991–2000	0.667	0.664	0.690	0.587 (0.689)	0.585	0.583	0.473	0.455
FRDC 2001–2010	0.671	0.667	0.692	0.688	0.598 (0.701)	0.597	0.474	0.452
FRDC 2011–2016	0.671	0.666	0.688	0.686	0.699	0.603 (0.705)	0.474	0.453
HOL	0.563	0.564	0.564	0.564	0.566	0.566	0.593 (0.661)	0.479
OTHER	0.544	0.544	0.544	0.545	0.544	0.545	0.552	0.740 (0.797)

The diagonals include diagonals (i.e., self-relationships) of $\boldsymbol{\Gamma}_8$(in parentheses) and $\boldsymbol{\Gamma}_{8MAF}$. Finnish Red dairy cattle (FRDC) have been divided into metafounders (MFs) by birth year, and Holstein and Other breeds have one MF per breed (HOL and OTHER, respectively).

Constructing $\mathbf{A_{22}}$ using $\boldsymbol{\Gamma}_8$ and $\boldsymbol{\Gamma}_{8MAF}$ increased the correlation between the diagonal elements of $\mathbf{G_{05}}$ and $\mathbf{A_{22}}$ from 0.66 to 0.76 (Table 7). The correlation between the diagonal elements of $\mathbf{A_{22}^{\Gamma_{8MAF}}}$ and $\mathbf{A_{22}}$ was higher (0.84) than that between $\mathbf{A_{22}^{\Gamma_8}}$ and $\mathbf{A_{22}}$ (0.81). The correlation between the diagonal elements of $\mathbf{G_{PvR1}}$ and $\mathbf{A_{22}}$ decreased from 0.53 to 0.33 and 0.37 for $\mathbf{A_{22}^{\Gamma_8}}$ and $\mathbf{A_{22}^{\Gamma_{8MAF}}}$, respectively (Table 7; III).

Table 7. Correlation of the diagonal (upper triangle) and off-diagonal (lower triangle) elements of $\mathbf{A_{22}}$, $\mathbf{A_{22}^{\Gamma_8}}$, $\mathbf{A_{22}^{\Gamma_{8MAF}}}$, $\mathbf{G_{PvR1}}$, and $\mathbf{G_{05}}$.

Matrix[1]	$\mathbf{A_{22}}$	$\mathbf{A_{22}^{\Gamma_8}}$	$\mathbf{A_{22}^{\Gamma_{8MAF}}}$	$\mathbf{G_{PvR1}}$	$\mathbf{G_{05}}$
$\mathbf{A_{22}}$	1	0.81	0.84	0.53	0.66
$\mathbf{A_{22}^{\Gamma_8}}$	0.89	1	0.99	0.33	0.76
$\mathbf{A_{22}^{\Gamma_{8MAF}}}$	0.92	0.99	1	0.37	0.76
$\mathbf{G_{PvR1}}$	0.89	0.86	0.88	1	0.70
$\mathbf{G_{05}}$	0.83	0.91	0.91	0.88	1

[1] $\mathbf{A_{22}}$ - the pedigree relationship matrix of genotyped animals; $\mathbf{A_{22}^{\Gamma_8}}$ and $\mathbf{A_{22}^{\Gamma_{8MAF}}}$ - the pedigree relationship matrices of genotyped animals augmented by the $\mathbf{\Gamma_8}$ and $\mathbf{\Gamma_{8MAF}}$; $\mathbf{G_{PvR1}}$ - the genomic relationship matrix constructed using the VanRaden (2008) method 1; $\mathbf{G_{05}}$ - the genomic relationship matrix with allele frequencies equal to 0.5.

4.3.2 Genomic estimated breeding values and validation of the model fit

The calculated correction factors $k_{\mathbf{\Gamma_8}}$ and $k_{\mathbf{\Gamma_{8MAF}}}$ used to adjust the genetic variance in the base population descending from the metafounders were 0.72 and 0.77, respectively. Average bull and cow GEBV 305d milk yield trends had a similar shape (III). Correlation of bull GEBVs between the MF model and the original 236 UPG model was higher (0.972) than the correlation between the MF model and the 8 UPG model (0.931). The GEBV standard deviation level for bulls born in 2012–2014 was 20 kg (3%) higher in the MF models than in the UPG models. Regression coefficients (b_1) were generally slightly higher and closer to 1 in MF models than in UPG models (Table 8).

In the bull validation set, similar adjusted model reliability values were obtained by $ssGBLUP_{8UPG}$, $ssGBLUP_{\Gamma_8}$, and $ssGBLUP_{\Gamma_{8MAF}}$, and the gain was 0.04 in comparison with $ssGBLUP_{236UPG}$. In the cow validation set, the validation reliability values using the MF models were higher by 0.01 than those achieved by the UPG models.

Table 8. Regression analysis results from the four single-step genomic best linear unbiased prediction (ssGBLUP) models in Finnish Red dairy cattle (FRDC) population

Validation set	Model	b_0	b_1	$R^2_{EDC/ERC}$
Bulls (101 animals)	$ssGBLUP_{236UPG}$	70	0.61	0.27
	$ssGBLUP_{8UPG}$	18	0.73	0.31
	$ssGBLUP_{r_8}$	-22	0.72	0.31
	$ssGBLUP_{r_{8MAF}}$	-27	0.73	0.31
Cows (3,551 animals)	$ssGBLUP_{236UPG}$	118	0.89	0.36
	$ssGBLUP_{8UPG}$	150	0.89	0.36
	$ssGBLUP_{r_8}$	12	0.90	0.37
	$ssGBLUP_{r_{8MAF}}$	-0.2	0.93	0.37

b_0 - general mean; b_1 - regression coefficient (multiplied by 2 in bulls because DYD represents half of the breeding value of the sire); $R^2_{EDC/ERC}$ - the coefficient of determination adjusted by the average reliability of phenotypes in the validation group

5 Discussion

5.1 Leningrad region genetic and genomic evaluations

The estimated variance components (I) suggested lower heritability values than those reported for large HOL populations. This can be most likely explained by excessive environmental variance and high residual variance, rather than by small additive variance (Boldman & Freeman, 1988). Slight improvement was observed after introducing the herd by sire interaction effect (II). However, the estimated heritability (0.21) was lower than that reported for HOL 305d data in the USA (0.29; Carabano et al. 1989), Japan (0.30; Suzuki et al., 1994), or Kenya (0.29; Ojango & Pollott, 2001).

The overall genetic trend for milk, fat, and protein yield in cows was positive but lower than that in countries from where the Holstein breed bulls had been imported (I, II). For instance, in Canada, the average annual EBV milk yield change in 2004-2014 was 85.4 kg (https://www.cdn.ca), whereas in the Leningrad region, it was 61.4 kg. Similar but less dynamic trends than those in Canada were observed for fat (1.90 vs 4.2 kg) and protein (1.76 vs 3.3 kg) yield. A positive genetic trend was most likely explained by on-farm management and importation of top sires (Kudinov et al., 2017).

The LR set of genotyped animals consisted of a limited number of progeny-tested bulls (301) and the number of genotyped cows with performance records was approximately thrice the number of bulls (893). Supplementation of the data by adding genomic data of DFS bulls that have genetic ties in both data sets did not sufficiently increase the size of the reference population (II). However, in bull validation tests for milk and fat yield, the regression coefficient $b_1$was slightly closer to 1 after adding the DFS genotypes. However, the validation reliability for milk yield did not change and that for fat yield did not increase sufficiently (by 0.01).

The obtained correlation of unweighted DYDs and GEBVs in milk using the LR data and genotypes was higher (0.38) than that reported by Ma et al. (2014) for a relatively small Chinese HOL reference population (85 bulls and 2,862 cows). The authors reported a correlation coefficient of 0.26 between DRP and the direct genomic breeding value for milk yield.

The highest validation reliability for milk yield was obtained with a model where DRP phenotypes of DFS bulls derived from MACE Interbull were combined with LR data (II). Pribyl et al. (2013) reported higher R^2 (0.67) obtained after incorporating Interbull DRP values into Czech HOL genomic values. The favorable effect of including the DFS data on R^2 in milk yield was not observed in fat yield. Two explanations could be proposed for this discrepancy. Firstly, the phenotypic data recordings had some limitations (Kudinov et al., 2017; I). Secondly, the commonly used Interbull validation practice (Mäntysaari et al. 2010) is not ideal for single-step models. As shown by Legarra & Reverter (2018), reciprocal of the size of contemporary groups may generate an upwards bias in the R^2 due small size of contemporary groups.

The shortcoming of ssGBLUP originates from the inconsistency in the definition of the base population in genotyped and non-genotyped animals (Misztal et al., 2013; Legarra et al., 2015). The LR population pedigree is characterized by a substantial number of ancestors originating from external populations; thus it is close to UPG in the pedigree. In the analysis of such pedigree, the MF is a promising way to solve the incompatibility in the pedigree and genomic relationship matrices. The MF approach was tested on the LR data but failed to improve the validation reliability (II). An explanation for the poor MF approach performance can be due to the small number of genotyped animals of which most were born during the last two decades. Proper estimation of base population AFs was difficult for HOL groups originating from different countries.

The genetic and further genomic evaluation were the main goals of the research (I, II) undertaken. However, an important outcome was to draw attention of breeders, farmers, and AI stations to accurate methods of breeding value prediction for local animals and to illustrate the importance of reliable data recording. The results of the integration of HOL DFS genomic data and MACE EBVs (II) in ssLR present the possibility of collaboration between DFS breeding program and LR farmers.

5.2 Metafounder approach in dairy cattle evaluations

5.2.1 Base population

Defining the base population is the greatest challenge in the MF approach. Two crucial issues should be kept in mind when MFs are designed: AF change over time and the linkage between the genotyped animals and the base population. Dairy cattle routine evaluation pedigree may have many UPGs that cannot be simply treated as MFs. At the same time, a small number of MFs should accurately depict the genetic change of a population over time. In the FRDC study, the genetic change was accounted for by using multiple MFs for the FRDC breed. The pedigree-based method, used to estimate AFs (Garcia-Baccino et al., 2017) assumes that the MFs are defined through a pedigree. In the data, a major part of the genotyped animals (75%) contributed to the oldest FRDC group (FRDC < 1971) when the full pedigree was used; however, most of the genotyped animals (90.6%) were born after 2000 (III). The issue of unbalanced distribution was solved by AF estimation using a limited pedigree where non-genotyped animals were included with genotyped offspring only. Applicability of the chosen pedigree limitation approach was confirmed by similarity of the average diagonal values of HOL MF in the **Γ** matrix estimated using FRDC (0.593) or HOL (0.615) genotypes.

5.2.2 Gamma matrix

The original MF publication (Legarra et al., 2015) presented the gamma matrix computed using AFs derived from VanRaden et al. (2011), which included SNP markers with MAF ≥5%. No certain rule on how to filter SNP markers to be used for **Γ** matrix estimation was presented. Theoretically, exclusion of low MAF markers will omit those with highly uncertain or erroneous AF estimates. Because the **Γ** matrix is a function of the chosen MAF threshold used in the marker selection, off-diagonal and diagonal values in $\mathbf{\Gamma_{8MAF}}$ were lower than those in $\mathbf{\Gamma_8}$ (III). The average diagonal values in $\mathbf{A_{22}^{\Gamma_{8MAF}}}$ were lower than those in $\mathbf{G_{05}}$, whereas the average diagonal values in $\mathbf{A_{22}^{\Gamma_8}}$ were higher than those in $\mathbf{G_{05}}$. However, selecting a MAF threshold >5% may cause unwanted similarity

between AFs of different MFs, and consequently, **Γ** matrix would have inflated (high) covariances between breeds.

Designed **Γ** matrices had >0 off-diagonal elements, suggesting shared genetics between breeds. The estimated average mean relationship of RDC and HOL breeds (0.47) in $\mathbf{\Gamma}_{8MAF}$ was close to those reported (Legarra et al., 2015) between Jersey breed and HOL (0.48). Use of the **Γ** matrix to construct the pedigree-based relationship matrix $\mathbf{A}_{22}^{\mathbf{\Gamma_8}}$ or $\mathbf{A}_{22}^{\mathbf{\Gamma_{8MAF}}}$ increased the correlation between elements of the pedigree and genomic relationship matrices when compared with the correlation between traditionally formed matrices ($\mathbf{G_{PvR1}}$ and $\mathbf{A_{22}}$).

5.2.3 Single-step evaluations

Genetic trends in GEBV from the UPG and MF models had a similar shape, indicating no effect of the alternative group or founder definitions. Expected reduction of the genetic trend owing to inadequate definition of groups (Tsuruta et al., 2014) was not observed (III). Perhaps, ssGBLUP predictions where most of the sires were genotyped were robust against the definition of UPG or MF. Meyer & Tier (2018) reported a slightly higher estimated genetic trend with the MF approach than with ssGBLUP without groups. However, females were the most often genotyped group in their data. The unstandardized genetic levels in the MF models were higher than those in the UPG models. This difference did not affect the animal rankings by GEBV but indicates that the models defined base population differently. Accuracy of genomic prediction was higher in MF models than in the original 236 UPG model. A reason for that can be incomplete QP (Quaas & Pollak, 1981) transformation used for non-purebred population (Bradford et al., 2019). Accuracy may be improved by using QP transformation in $\mathbf{H^{-1}}$ (Misztal et al., 2013) for the 236 UPG model, as was done using LR data (II).

5.3 Implications and future developments

5.3.1 LR dairy breeding

Results from this thesis can be used by the LR dairy industry to update the genetic evaluation system and move toward a routine genomic prediction. The data editing practices developed and results from the BLUP-AM model application were presented to LR farmers and breeding committees. The observed data recording pitfalls and peculiarities found were reported to the local recording centre Plinor LLC. A sampling of cows for genotyping was undertaken in a way to include most fastidious and progressive farmers first. This induced interest not only in those farmers but in the whole farmer community. The MME (I) and data editing practices developed in this study were used as the starting point in research and development work in LR Ayrshire cattle. Despite many challenges in the Russian dairy sector, the use of modern evaluation methods is a good start for the industry to change the current working strategy. Current selection of breeding animals in LR is based on single traits or foreign indexes for imported bulls. In the future, own production index should be developed.

Extensive outreach work should be continued to get farmers more involved in setting up their own genomic evaluation system. It would be beneficial for LR to become an ICAR member and follow the recording guidelines and also become part of Interbull MACE evaluations. Reliable recording would also help to create joint evaluation with DFS in a manner similar to EuroGenomics cooperation (https://www.eurogenomics.com/). The LR reference population should be expanded by continuous genotyping of the new progeny-tested bulls and cows. Cow genotyping should be performed to encompass as many herds as possible. The natural future direction to improve the prediction accuracy in the LR data is to use test-day records instead of 305d data. Development of genetic evaluations for nonproductive traits (fertility, health, and welfare) should be started.

5.3.2 Metafounder approach

An ssGBLUP is seldom used to perform routine genomic predictions in dairy cattle. The interest of the industry to use the method is constantly growing. The MF approach is one of the most promising approaches to resolve the compatibility issues. This thesis presented results on **Γ** matrix construction considered in continuous single-step research work performed in Nordic countries. Further research on the MF approach is needed. For instance, it was noticed (III) that the behavior of genomic prediction was not considerably changed when the off-diagonal elements of the **Γ** matrix were reduced. Further, replacement of many UPGs in routine dairy cattle pedigree by few MFs may generate unreliable results. This was not the case in small FRDC data but expected in the large Nordic RDC or HOL population. The limited number of MF did not present genetic base differences as detailed as the many UPGs. The Nordic HOL and RDC evaluations are actively testing approaches to extend the number of MFs to the number of UPGs using covariance functions.

6 Conclusions

This thesis demonstrates the ways to implement single-step genomic prediction in small-scale cattle populations. The ssGBLUP prediction was performed for the LR RBW and HOL breeds and FRDC population.

Data editing practices were developed to prepare the LR data to be used in BLUP-AM and further in ssGBLUP prediction (I, II). Two BLUP MME were proposed (I, II). According to the results, inclusion of herd by sire interaction random effect into the model provided a better data fit and better accounted for environmental differences between herds (II).

The group of genotyped animals used to implement LR ssGBLUP was small and had more cows than bulls (II). Genomic prediction model was implemented, but its prediction accuracy was low. The reliability of ssGBLUP prediction was enhanced by the integration of external DFS information (II). Limited reliability improvement was shown for the fat trait presumably originating from data recording. The results suggest that the collaboration of LR farmers with the DFS breeding program could serve as way towards a routine genomic evaluation.

The MF approach was implemented on complicated multi-breed pedigree of FRDC (III). The approach developed for pedigree truncation and inclusion of MAF filtering resulted in the $\boldsymbol{\Gamma}$ matrix ($\boldsymbol{\Gamma}_{8\mathrm{MAF}}$), which perfectly fit the FRDC pedigree. The use of MFs in ssGBLUP increased the validation reliability and decreased bias in comparison with the original UPG model.

During this study, educational activities based on the results were widely undertaken for LR farmers and breeders to ensure immediate impact of the research.

7 Literature

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S. & Lawlor, T.J. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science, 93:743-752.

Arzumanyan, E.A., Markin, E.F. & Ryabov, U.K. 1973. Ural Black and White cattle, Moscow, Kolos publish. 20 p. (in Russian).

Boldman, K.G. & Freeman, A.E. 1988. Estimates of Genetic and Environmental Variances of First and Later Lactations at Different Production Levels. Journal of Dairy Science, 71: 81–82.

Bradford, H. L., Masuda, Y., VanRaden, P.M., Legarra, A. & Misztal, I. 2019. Modelling missing pedigree in single-step genomic BLUP. Journal of Dairy Science, 102:2336–2346.

Carabaño, M.J., Van Vleck, L.D., Wiggans, G.R. & Alenda, R. 1989. Estimation of Genetic Parameters for Milk and Fat Yields of Dairy Cattle in Spain and the United States, Journal of Dairy Science, 72:3013-3022.

Christensen, O.F. & Lund, M.S. 2010. Genomic prediction when some animals are not genotyped. Genetic Selection Evolution, 42:2.

Christensen, O.F., Madsen, P., Nielsen, B., Ostersen, T. & Su, G. 2012. Single-step methods for genomic evaluation in pigs. Animal, 6:1565-1571.

Ding, X., Zhang, Z., Li, X., Wang, S., Wu, X., Sun, D., Yu, Y., Liu, J., Wang, Y., Zhang, Y., Zhang, S., Zhang, Y. & Zhang, Q. 2013. Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows. Journal of Dairy Science, 96:5315-5323.

Garcia-Baccino, C. A., Legarra A., Christensen, O. F., Misztal, I, Pocrnic, I., Vitezica, Z.G. & Cantet, R.J. 2017. Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. Genetic Selection Evolution, 49:34.

Goddard, M.E. & Hayes, B.J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics, 10:381–391.

Guarini, A.R., Lourenco, D.A.L., Brito, L.F., Sargolzaei, M., Baes, C.F., Miglior, F., Tsuruta, S. & Misztal, I. 2019. Use of a single-step approach for integrating foreign information into national genomic evaluation in Holstein cattle. Journal of Dairy Science, 102:8175-8183.

Ignashkina, A.A. & Kuznetsov, V.M. 1988. Breeding value evaluation of bulls using MCC and BLUP methods. RRIFAGB Bulletin 101:3–5. (in Russian).

Instruction for the inspection and evaluation of bulls of dairy and meat milk breeds on the quality of offspring. 1979. Moscow, USSR Ministry of Agriculture (in Russian).

Jairath, L., Dekkers, J.C., Schaeffer, L.R., Liu, Z., Burnside, E.B. & Kolstad, B. 1998. Genetic evaluation for herd life in Canada. Journal of Dairy Science, 81:550-562.

Jorjani, H., Jakobsen, J., Hjerpe, E., Palucci, V. & Dürr, J. 2012. Status of genomic evaluation in the Brown Swiss populations. Interbull Bulletin, 46:46-54.

Kudinov, A.A., Juga, J., Uimari, P., Mäntysaari, E.A., Strandén, I., Plemyashov, K.V., Saksa, E.I. & Smaragdov, M.G. 2017. Upgrading Dairy Cattle Evaluation System in Russian Federation. Interbull Bulletin, 51: 67–74.

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I. & Misztal, I. 2015. Ancestral relationships using metafounders: Finite ancestral populations and across population relationships. Genetics, 200:455-468.

Legarra, A. & Reverter, A. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Genetic Selection Evolution, 50: 53.

Li, X., Wang, S., Huang, J., Li, L., Zhang, Q. & Ding, X. 2014. Improving the accuracy of genomic prediction in Chinese Holstein cattle by using one-step blending. Genetic Selection Evolution, 46:66.

Lund, M.S., de Roos, A.P.W., de Vries, A.G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F. & Su, G. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genetic Selection Evolution, 43:43.

Ma, P., Lund, M.S., Ding, X., Zhang, Q. & Su, G. 2014. Increasing imputation and prediction accuracy for Chinese Holsteins using joint Chinese-Nordic reference population. Journal of Animal Breeding and Genetics. 131:462-472.

Madsen, P., Su, G., Labouriau, R. & Christensen, O.F. 2010. DMU - a package for analyzing multivariate mixed models. I CD communication. Proceeding of the 9th WCGALP, Leipzig, Germany, August 1-6, 732.

McPeek, M. S., Xiaodong W., & Ober, C. 2004. Best Linear Unbiased Allele-Frequency Estimation in Complex Pedigrees. Biometrics 60:359–367.

Meyer, K., Tier B. & Swan, A. 2018. Estimates of genetic trend for single-step genomic evaluations. Genetic Selection Evolution, 50:39.

Meuwissen, T.H., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157:1819-1829.

Misztal, I., Aguilar, I., Legarra, A. & Lawlor, T.J. 2010. Choice of parameters for single-step genomic evaluation for type. Journal of Dairy Science,93:533.

Misztal, I., Vitezica, Z.G., Legarra, A., Aguilar, I., & Swan, A.A. 2013. Unknown-parent groups in single-step genomic evaluation. Journal of Animal Breeding and Genetics, 130:252–258.

Myakoshina, L.A., Klimez, N.V. & Kuznetsov, V.M. 1992. BLUP models choosing for bulls evaluation. RRIFAGB Bulletin, 133: 12–15. (in Russian).

Mäntysaari, E.A., Liu, Z. & VanRaden, P.M. 2010. Interbull validation test for genomic evaluations. Interbull Bulletin, 41:17-22.

Mäntysaari, E.A., Koivula, M. & Strandén, I. 2020. Symposium review: Single-step genomic evaluations in dairy cattle. Journal of Dairy Science, 103:5314-5326.

Ojango, J. & Pollott, G. 2001. Genetics of milk yield and fertility traits in Holstein-Friesian cattle on large-scale Kenyan farms. Journal of Animal Science, 79: 1742–1750.

Patterson, H.D. & Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika, 58:545-554.

Přibyl, J., Madsen, P., Bauer, J., Přibylová, J., Šimečková, M., Vostrý, L. & Zavadilová, L. 2013. Contribution of domestic production records, Interbull estimated breeding values, and single nucleotide polymorphism genetic markers to the single-step genomic evaluation of milk production. Journal of Dairy Science, 96:1865-1877.

Quaas, R.L. & Pollak, E.J. 1981. Modified equations for sire models with groups. Journal of Dairy Science, 64:1868–1872.

Schaefer, L.R. 2013. History of genetic evaluation methods in dairy cattle.

Shkirando, U.P. 1986. Increasing the effectiveness of evaluation of the genotype of dairy cattle using Breeding Value Indexes, list squares and BLUP. Dissertation. (in Russian).

Smaragdov, M.G., Kudinov, A.A. & Uimari, P. 2018. The Assessing the genetic differentiation of Holstein cattle herds in the Leningrad region using Fst statistics. Agricultural and Food Science, 27:96–101.

Song, H., Zhang, J., Zhang, Q. & Ding, X. 2019. Using Different Single-Step Strategies to Improve the Efficiency of Genomic Prediction on Body Measurement Traits in Pig. Frontiers in genetics,14:730.

Sponenberg, D.P. & Bixby, D.E. 2007. Managing Breeds for a Secure Future: Strategies for Breeders and Breed Associations. American Livestock Breeds Conservancy. Pittsboro, NC.

Strandén, I. & Lidauer, M. 1999. Solving large mixed models using preconditioned conjugate gradient iteration. Journal of Dairy Science, 82:2779–2787.

Strandén, I. & Vuori, K. 2006. RelaX2: pedigree analysis program. In: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production: 13-18 August 2006; Belo Horizonte, MG, Brazil, 27-30.

Strandén, I. & Mäntysaari, E. A. 2010. A recipe for multiple trait deregression. Interbull Bulletin, 42:21-24.

Strandén, I. & Mäntysaari, E.A. 2020. Bpop: an efficient program for estimating base population allele frequencies in single and multiple group structured populations. Agricultural and Food Science, 29:166–176.

Strandén, I. & Mäntysaari, E.A. 2018. HGinv Program. Natural Resources Institute Finland (LUKE).

Suzuki, M. & Van Vleck, L.D. 1994. Heritability and Repeatability for Milk Production Traits of Japanese Holsteins from an Animal Model. Journal of Dairy Science, 77:2.

Taskinen, M., Mäntysaari, E.A., Aamand, G.P. & Strandén., I. 2014. Comparison of breeding values from single-step and bivariate blending methods. In Proceedings of the 10th World Congress on Genetics Applied to Livestock Production: August 2014, Vancouver, BC, Canada, 17-22.

Tsuruta, S., Misztal, I., Lourenco, D. & Lawlor T. 2014. Assigning unknown parent groups to reduce bias in genomic evaluations of final score in US Holsteins. Journal of Dairy Science, 97:5814-5821.

Vandenplas, J., Colinet, F.G. & Gengler, N. 2014. Unified method to integrate and blend several, potentially related, sources of information for genetic evaluation. Genetics Selection Evolution, 46:59.

VanRaden, P.M. 2001. Methods to combine estimated breeding values obtained from separate sources. Journal of Dairy Science, 84:47-55.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions, Journal of Dairy Science, 91:4414-4423.

VanRaden, P.M. 2012. Avoiding bias from genomic pre-selection in converting daughter information across countries. Interbull Bulletin, 45.

VanRaden, P.M. 2020. Symposium review: How to implement genomic selection. Journal of Dairy Science, 103:5291-5301.

Vitezica, Z., Aguilar, I., Misztal, I. & Legarra, A. 2011. Bias in genomic predictions for populations under selection. Genetic Resources, 93:357–366.

Wiggans, G.R, Cole, J.B., Hubbard, S.M. & Sonstegard, T.S. 2017. Genomic Selection in Dairy Cattle: The USDA Experience. Annual Review of Animal Biosciences, 5:309-327.

Xiang, T., Christensen, O.F. & Legarra, A. 2017. Genomic evaluation for crossbred performance in a single-step approach with metafounders. Journal of Animal Science, 95: 1472–1480.

Yearbook of breeding work in dairy cattle of Russian Federation in 2016. 2017. Moscow, VNIIPlem (in Russian).